Sustainability of Digital Formats: Planning for Library of Congress Collections

XML (Extensible Markup Language)

>> Back

Table of Contents

Format Description Properties

- ID: fdd000075
- Short name: XML
- Content categories: text, dataset
- Format Category: file-format, encoding
- Other facets: text, structured, symbolic
- Last significant FDD update: 2022-04-14
- Draft status: Partial

Identification and description

Full name	Extensible Markup Language (XML)
Description	Extensible Markup Language (XML) is a simple, very flexible text format derived from SGML (ISO 8879). XML documents fall into two broad categories: data-centric and document-centric. Data-centric documents are those where XML is used as a data transport. Examples include sales orders, patient records, directory entries, and metadata records. One significant use of data-centric XML is for manifests (lists) of digital content; another is for metadata embedded into digital content files. Document-centric documents are those in which XML is used for its SGML-like capabilities, reflecting the structure of particular classes of documents, such as books with chapters, user manuals, newsfeeds and articles incorporating explicit metadata in addition to the text. An XML document's markup structure can be defined by a schema language and validated against a definition in that language. The initial, and as of 2008, most widely used schema languages are the Document

	Type Definition (DTD) language and W3C XML Schema. Other schema languages exist, including RDF and RELAX-NG.
Production phase	Can be used as initial, middle, or final-state format.
Relationship to other formats	
Has subtype	XML_1_0, XML (Extensible Markup Language) 1.0
Has subtype	XML_1_1, XML (Extensible Markup Language) 1.1
Has subtype	XML_DTD, Document Type Definition
Has subtype	XML_SCHEMA, W3C XML Schema Language
May contain	CSS, Cascading Style Sheet (CSS) Markup. May embed CSS markup directly or invoke an external CSS file.
Used by	IMF_Package, Interoperable Master Format (IMF). Used for mandatory Asset Map and Packing List in IMF Package
Used by	APK, Android Package
Used by	IPA, iOS App Store Package
Used by	XAP, Silverlight Application Package
Has modified version	Other entities have introduced variant format versions using the .zip extension, not strictly compatible with any particular chronological version of ZIP_PK, but using its extension capabilities. See Notes below for a brief discussion of variants and compatibility.
Has extension	ADM, Audio Definition Model
Has extension	PEF, Portable Embosser Format

Local use

LC experience or existing holdings	Used by LC to represent metadata records (including MARC bibliographic and authority records, MODS, METS) for web-compatible interchange, in particular using the Open Archives Initiative Protocol for Metadata Harvesting and SRU (Search/Retrieval via URL).
LC preference	The Library of Congress Recommended Formats Statement (RFS) lists XML as a Preferred format for Textual Works - Digital, with included or accessible DTD/schema, XSD/XSL presentation stylesheet(s), and explicitly stated character encoding. LC will express preferences based on specific DTDs, W3C XML Schema instances, or instance documents in other schema languages for defining XML-based formats. LC will prefer XML that represents the structure of documents rather than layout. The RFS also lists the XML format as an Acceptable format for Textual Works - Digital for XML-based document formats with presentation stylesheets. In addition to textual works in digital form, the Recommended Formats Statement lists XML as a Preferred format schema for Dataset metadata, for packaging data for Video - File-Based and Physical Media and for metadata for Audio Works - Media Independent (digital). The RFS also lists XML as an Acceptable format for metadata for photographs in digital form, other graphic images in digital form, and 2D and 3D Computer Aided Design vector images and scanned 3D objects (output from photogrammetry scanning).

Sustainability factors

Disclosure	Open standard. Developed by W3C (World Wide Web Consortium). To be useful for interoperability or long-term content preservation, an XML document must be associated with a schema specification for the elements and tags it contains. Such schema specifications (see XML_DTD and XML_XSD) must also be disclosed.
Documentation	Maintained by W3C [http://www.w3.org/XML/]. Specifications for the two versions as of 2008 are at Extensible Markup Language (XML) 1.0 and Extensible Markup Language (XML) 1.1.
Adoption	Very widely adopted as the basis for interchange of documents and data over the Web. Many generic tools exist, including free and open source software. Major software vendors have all incorporated support for XML in some form.
Licensing and patents	None
Transparency	XML is human-readable and designed for straightforward automatic parsing. For the contents to be understood, a well-documented DTD, XML Schema, or other specification is needed. Human-comprehensible element tags are advantageous for transparency.
Self-documentation	XML is widely used as a syntax for metadata, and metadata for all purposes can be embedded in XML documents with appropriate schema specifications. **Accessibility Features** XML-based formats have good support for accessibility features. According to W3C's XML Accessibility Guidelines, XML-based formats can include features that promote accessibility depending on implementation. This document outlines some techniques to achieve this, including the following: • Ensure that authors can associate multiple media objects as alternatives for any content (including images, movies, songs, etc.). (1.1) • Ensure all semantics are captured in markup in a repurposable form, so that they are presented in a device independent way. (2.1) • Separate presentation properties using stylesheet technology/styling mechanisms, which allows content to adapt to presentational needs for the user such as larger font or more contrast. (2.2) • Use the standard XML linking and pointing mechanisms (XLink and XPointer). (2.3) • Define element types that allow classification and grouping (header, section, list, etc.). (2.4) • Provide a mechanism for identifying summary/abstract/title. (2.7) • Define navigable structures that allow discrete, sequential, structured, and search navigation functionalities. (3.2)
External dependencies	None
Technical protection considerations	None

Quality and functionality factors

Text	
Normal rendering	XML can represent all UNICODE characters, with UTF-8 being the default character encoding. XML tagging offers potential for explicitly representing logical structure of text, such as paragraphs and headings, and character emphasis (bold, italics, etc.). Effective support for normal rendering is dependent on an appropriate DTD or schema specification.
Integrity of document structure	XML is ideal for representing document structure.
Integrity of layout and display	For textual content, best practice is to have the XML represent the logical document structure and use stylesheets to render the text in a form appropriate for the end user.
Support for mathematics, formulae, etc.	Requires specialized markup (e.g., MathML) and corresponding rendering engine. Scholars in many scientific disciplines are not satisfied with the performance of such rendering engines.
Functionality beyond normal rendering	Depends on particular DTD or schema specification.

File type signifiers and format identifiers

Tag	Value	Note
Filename extension	xml	Common practice for XML document instances is to use the .xml extension. The particular schema or DTD should be declared within the document. Some schemas specify the use of different file extensions.
Internet Media Type	text/xml application/xml	If an XML document is readable by casual users, *text/xml* is preferred. See RFC 3023 for further details.
Magic numbers	See note.	Although no byte sequences can be counted on to always be present, XML MIME entities in ASCII-compatible charsets (including UTF-8) often begin with hexadecimal 3C 3F 78 6D 6C ("<?xml"), and those in UTF-16 often begin with hexadecimal FE FF 00 3C 00 3F 00 78 00 6D 00 6C or FF FE 3C 00 3F 00 78 00 6D 00 6C 00 (the Byte Order Mark (BOM) followed by "<?xml"). See RFC 3023 for further details.
Pronom PUID	fmt/101	See http://www.nationalarchives.gov.uk/PRONOM/fmt/101.
Wikidata Title ID	Q2115	See https://www.wikidata.org/wiki/Q2115.
Other	NF00654	See https://www.archives.gov/files/lod/dpframework/id/NF00654.ttl.

Notes

General	The original design goals for XML were: • 1. XML shall be straightforwardly usable over the Internet.

	• 2. XML shall support a wide variety of applications. • 3. XML shall be compatible with SGML. • 4. It shall be easy to write programs which process XML documents. • 5. The number of optional features in XML is to be kept to the absolute minimum, ideally zero. • 6. XML documents should be human-legible and reasonably clear. • 7. The XML design should be prepared quickly. • 8. The design of XML shall be formal and concise. • 9. XML documents shall be easy to create. • 10. Terseness in XML markup is of minimal importance. Style sheets, for example in CSS or XSLT, can be associated with XML documents for presentation of XML files on the web. See Associating Style Sheets with XML documents, which includes an example with <?xml-stylesheet> processing instructions for associating external CSS style sheets with an XML file. The XHTML format specification also includes a <link> element that can be used to invoke external style sheets.
History	"XML is primarily intended to meet the requirements of large-scale Web content providers for industry-specific markup, vendor-neutral data exchange, media-independent publishing, one-on-one marketing, workflow management in collaborative authoring environments, and the processing of Web documents by intelligent clients. It is also expected to find use in certain metadata applications. XML is fully internationalized for both European and Asian languages, with all conforming processors required to support the Unicode character set in both its UTF-8 and UTF-16 encodings. The language is designed for the quickest possible client-side processing consistent with its primary purpose as an electronic publishing and data interchange format." [from 1997-12-08 W3C press release] See http://www.w3.org/XML/hist2002.

Format specifications

- Latest specifications as of March 2008.
  - Extensible Markup Language (XML) 1.0 (Fourth Edition) (http://www.w3.org/TR/xml/). W3C Recommendation 16 August 2006, Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, François Yergeau eds.
  - Extensible Markup Language (XML) 1.1 (Second Edition) (http://www.w3.org/TR/xml11/). W3C Recommendation, 16 August 2006, Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, François Yergeau, John Cowan, ed.

Useful references

URLs

- Extensible Markup Language (XML) activity at W3C (http://www.w3.org/XML/).
- XML Development History (http://www.w3.org/XML/hist2002). From W3C. Covers 1996-2000.
- XML Media Types (http://www.ietf.org/rfc/rfc3023.txt). Registration of Media Types with IANA. January 2001.

- PRONOM entry for fmt/101 (http://www.nationalarchives.gov.uk/PRONOM/fmt/101). Information in PRONOM from UK National Archives about Extensible Markup Language 1.0. PUID: fmt/101.
- Wikidata entry for Q2115 (https://www.wikidata.org/wiki/Q2115). Information in Wikidata about Extensible Markup Language. Wikidata Title ID: Q2115
- NARA File Format Preservation Plan ID entry for NF00654 (https://www.archives.gov/files/lod/dpframework/id/NF00654.ttl). Information in NARA File Format Preservation Plan ID about Extensible Markup Language (XML).

Last Updated: 05/08/2024

Digital Preservation Home | Digital Formats Home